Regularization Techniques for Learning with Matrices

Sham M. Kakade
The Wharton School
University of Pennsylvania
skakade@wharton.upenn.edu

Shai Shalev-Shwartz
School of Computer Sc. & Eng.
The Hebrew University of Jerusalem
shais@cs.huji.ac.il

Ambuj Tewari
Computer Science Department
University of Texas at Austin
ambuj@cs.utexas.edu

Abstract

There is growing body of learning problems for which it is natural to organize the parameters into matrix, so as to appropriately regularize the parameters under some matrix norm (in order to impose some more sophisticated prior knowledge). This work describes and analyzes a systematic method for constructing such matrix-based, regularization methods. In particular, we focus on how the underlying statistical properties of a given problem can help us decide which regularization function is appropriate.

Our methodology is based on the known duality fact: that a function is strongly convex with respect to some norm if and only if its conjugate function is strongly smooth with respect to the dual norm. This result has already been found to be a key component in deriving and analyzing several learning algorithms. We demonstrate the potential of this framework by deriving novel generalization and regret bounds for multi-task learning, multi-class learning, and kernel learning.

1 Introduction

As we tackle more challenging learning problems, there is an increasing need for algorithms that efficiently impose more sophisticated forms of prior knowledge. Examples include: the group Lasso problem (for "shared" feature selection across problems), kernel learning, multi-class prediction, and multi-task learning. A central question here is to understand the performance of such algorithms in terms of the attendant complexity restrictions imposed by the algorithm. Such analyses often illuminate the nature in which our prior knowledge is being imposed.

The predominant modern method for imposing complexity restrictions is through regularizing a vector of parameters, and much work has gone into understanding the relationship between the nature of the regularization and the implicit prior knowledge imposed, particular for the case of regularization with ℓ_2 and ℓ_1 norms (where one is more tailored to rotational invariance and margins, while the other is more suited to sparsity). When dealing with more complex problems, we need systematic tools for designing more complicated regularization schemes. This work examines regularization based on group norms and spectral norms of *matrices*. We analyze the performance of such regularization methods and provide a methodology for choosing a regularization function based on the underlying statistical properties of a given problem.

In particular, we utilize a recently developed methodology, based on the notion of *strong* convexity, for designing and analyzing the regret or generalization ability of a wide range of learning algorithms (see e.g. Shalev-Shwartz [2007], Kakade et al. [2008]). In fact, most of our efficient algorithms (both in the batch and online settings) impose some complexity control via the use of some *strictly* convex penalty function either explicitly via a regularizer or implicitly in the design of an online update rule. Central to understanding these algorithms is the manner in which these penalty functions are strictly convex, i.e. the behavior of the "gap" by which these convex functions lie above their tangent planes, which is strictly positive for strictly convex functions. Here, the notion of strong convexity provides one means to characterize this gap in terms of some general norm rather than just Euclidean.

The importance of strong convexity can be understood using the duality between strong convexity and strong smoothness. Strong smoothness measures how well a function is approximated at some point by its linearization. Linear functions are easy to manipulate (e.g. because of the linearity of expectation). Hence, if a function is sufficiently smooth we can more easily control its behavior. We further distill the analysis given in Shalev-Shwartz [2007], Kakade et al. [2008] — based on the strong-convexity/smoothness duality, we derive a key inequality which seamlessly enables us to design and analyze a family of learning algorithms.

Our focus in this work is on learning with matrices. We characterize a number of matrix based regularization functions, of recent interest, as being strongly convex functions — allowing us to immediately derive learning algorithms by relying on the family of learning algorithms mentioned previously. Specifying the general performance bounds for the specific matrix based regularization method, we are able to systematically decide which regularization function is more appropriate based on underlying statistical properties of a given problem.

1.1 Our Contributions

We can summarize the contributions of this work as follows:

- We show how the framework based on strong convexity/strong smoothness duality (see Shalev-Shwartz [2007], Kakade et al. [2008]) provides a methodology for analyzing matrix based learning methods, which are of much recent interest. These results reinforce the usefulness of this framework in providing both learning algorithms, and their associated complexity analysis. For this reason, we further distill the analysis given in Shalev-Shwartz [2007], Kakade et al. [2008] by emphasizing a key inequality which immediately enables us to design and analyze a family of learning algorithms.

- We provide template algorithms (both in the online and batch settings) for a number of machine learning problems of recent interest, which use matrix parameters. In particular, we provide a simple derivation of generalization/mistake bounds for: (i) online and batch multi-task learning using group or spectral norms, (ii) online multi-class categorization using group or spectral norms, and (iii) kernel learning.

- Based on the derived bounds, we interpret how statistical properties of a given problem can help us decide which regularization function is appropriate. For example, for the case of multi-class learning, we describe and analyze a new "group Perceptron" algorithm and show that with a shared structure between classes, this algorithm significantly outperforms previously proposed algorithms. Similarly, for the case of multi-task learning, the pressing question is what shared structure between the tasks allows for sample complexity improvements and by how much? We discuss these issues based on our regret and generalization bounds.

- Our unified analysis significantly simplifies previous analyses of recently proposed algorithms. For example, the generality of this framework allows us to simplify the proofs of previously proposed regret bounds for online multi-task learning (e.g. Cavallanti et al. [2008], Agarwal et al. [2008]). Furthermore, bounds that follow immediately from our analysis are sometimes much sharper than previous results (e.g. we improve the bounds for multiple kernel learning given in Lanckriet et al. [2004], Srebro and Ben-David [2006]).

1.2 Related work

We first discuss related work on learning with matrix parameters then discuss the use of strong convexity in learning.

Matrix Learning: This is growing body of work studying learning problems in which the parameters can be organized as matrices. Several examples are multi-class categorization (e.g. Crammer and Singer [2000]), multi-task and multi-view learning (e.g. Cavallanti et al. [2008], Agarwal et al. [2008]), and online PCA [Warmuth and Kuzmin, 2006]. It was also studied under the framework of group Lasso (e.g. Yuan and Lin [2006], Obozinski et al. [2007], Bach [2008]).

In the context of learning vectors (rather than matrices), the study of the relative performance of different regularization techniques based on properties of a given task dates back to Littlestone [1988], Kivinen and Warmuth [1997]. In the context of batch learning, it was studied by several authors (e.g. Ng [2004]).

We also note that much of the work on multi-task learning for regression is on union support recovery — a setting where the generative model specifies a certain set of relevant features (over all the tasks), and the analysis here focuses on the conditions and sample sizes under which the union of the relevant features can be correctly identified (e.g. Obozinski et al. [2007], Lounici et al. [2009]). Essentially, this is a generalization of the issue of identifying the relevant feature set in the standard single task regression setting, under ℓ_1 regression. In contrast, our work focuses on the agnostic setting of just understanding the sample size needed to obtain a given error rate (rather than identifying the relevant features themselves).

We also discuss related work on kernel learning in Section 6. Our analysis here utilizes the equivalence between kernel learning and group Lasso (as noted in Bach [2008]).

Strong Convexity/Strong Smoothness: The notion of *strong convexity* takes its roots in optimization. Zalinescu [2002] attributes it to a paper of Polyak in the 1960s. Relatively recently, its use in machine

learning has been two fold: in deriving regret bounds for online algorithms and generalization bounds in batch settings.

The duality of strong convexity and strong smoothness was first used by Shalev-Shwartz and Singer [2006], Shalev-Shwartz [2007] in the context of deriving low regret online algorithms. Here, once we choose a particular strongly convex penalty function, we immediately have a family of algorithms along with a regret bound for these algorithms that is in terms of a certain strong convexity parameter. A variety of algorithms (and regret bounds) can be seen as special cases.

A similar technique, in which the Hessian is directly bounded, is described by Grove et al. [2001], Shalev-Shwartz and Singer [2007]. Another related approach involved bounding a Bregman divergence [Kivinen and Warmuth, 1997, 2001, Gentile, 2003] (see Cesa-Bianchi and Lugosi [2006] for a detailed survey). Another interesting application of the very same duality is for deriving and analyzing boosting algorithms [Shalev-Shwartz and Singer, 2008].

More recently, Kakade et al. [2008] showed how to use the very same duality for bounding the Rademacher complexity of classes of linear predictors. That the Rademacher complexity is closely related to Fenchel duality was shown in Meir and Zhang [2003], and the work in Kakade et al. [2008] made the further connection to strong convexity. Again, under this characterization, a number of generalization and margin bounds (for methods which use linear prediction) are immediate corollaries, as one only needs to specify the strong convexity parameter from which these bounds easily follow (see Kakade et al. [2008] for details).

The concept of strong smoothness (essentially a second order upper bound on a function) has also been in play in a different literature, for the analysis of the concentration of martingales in *smooth* Banach spaces [Pinelis, 1994, Pisier, 1975]. This body of work seeks to understand the concentration properties of a random variable $\|X_t\|$, where X_t is a (vector valued) martingale and $\|\cdot\|$ is a smooth norm, say an L_p-norm.

Recently, Juditsky and Nemirovski [2008] used the fact that a *norm* is strongly convex if and only if its conjugate is strongly smooth. This duality was useful in deriving concentration properties of a random variable $\|\mathbf{M}\|$, where now $\mathbf{M}$ is a random matrix. The norms considered here were the (Schatten) L_p-matrix norms and certain "block" composed norms (such as the $\|\cdot\|_{2,q}$ norm).

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we describe the general family of learning algorithms. In particular, after presenting the duality of strong-convexity/strong-smoothness, we isolate an important inequality (Corollary 4) and show that this inequality alone seamlessly yields regret bounds in the online learning model and Rademacher bounds (that leads to generalization bounds in the batch learning model). We further highlight the importance of strong convexity to matrix learning applications by drawing attention to families of strongly convex functions over matrices. To do so, we rely on the recent results of Juditsky and Nemirovski [2008]. In particular, we obtain a strongly convex function over matrices based on strongly convex vector functions, which leads to a number of corollaries relevant to problems of recent interest. Next, in Section 3 we show how the obtained bounds can be used for systematically choosing an adequate prior knowledge (i.e. regularization) based on properties of the given task. We then turn to describe the applicability of our approach to more complex prediction problems. In particular, we study multi-task learning (Section 4), multi-class categorization (Section 5), and kernel learning (Section 6). Naturally, many of the algorithms we derive have been proposed before. Nevertheless, our unified analysis enables us to simplify previous analyzes, understand the merits and pitfalls of different schemes, and even derive new algorithms/analyses.

2 Preliminaries and Techniques

In this section we describe the necessary background. Most of the results below are not new and are based on results from Shalev-Shwartz [2007], Kakade et al. [2008], Juditsky and Nemirovski [2008]. Nevertheless, we believe that the presentation given here is simpler and slightly more general.

Our results are based on basic notions from convex analysis and matrix computation. The reader not familiar with some of the objects described below may find short explanations in Appendix A.

2.1 Notation

We consider convex functions $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, where $\mathcal{X}$ is a Euclidean vector space equipped with an inner product $\langle \cdot, \cdot \rangle$. We denote $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$. The subdifferential of f at $x \in \mathcal{X}$ is denoted by $\partial f(x)$. The Fenchel conjugate of f is denoted by $f^\star$. Given a norm $\|\cdot\|$, its dual norm is denoted by $\|\cdot\|_\star$. We say that a convex function is V-Lipschitz w.r.t. a norm $\|\cdot\|$ if for all $x \in \mathcal{X}$ exists $v \in \partial f(x)$ with $\|v\| \leq V$. Of particular interest are p-norms, $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$.

When dealing with matrices, We consider the vector space $\mathcal{X} = \mathbb{R}^{m \times n}$ of real matrices of size $m \times n$ and the vector space $\mathcal{X} = \mathbb{S}^n$ of symmetric matrices of size $n \times n$, both equipped with the inner product, $\langle \mathbf{X}, \mathbf{Y} \rangle := \mathrm{Tr}(\mathbf{X}^\top \mathbf{Y})$. Given a matrix $\mathcal{X}$, the vector $\sigma(\mathbf{X})$ is the vector that contains the singular values of

$\mathbf{X}$ in a non-increasing order. For $\mathbf{X} \in \mathbb{S}^n$, the vector $\lambda(\mathbf{X})$ is the vector that contains the eigenvalues of $\mathbf{X}$ arranged in non-increasing order.

2.2 Strong Convexity–Strong Smoothness Duality

Recall that the domain of $f : \mathcal{X} \to \mathbb{R}^*$ is $\{x : f(x) < \infty\}$ (allowing f to take infinite values is the effective way to restrict its domain to a proper subset of $\mathcal{X}$). We first define strong convexity.

Definition 1 *A function $f : \mathcal{X} \to \mathbb{R}^*$ is β-strongly convex w.r.t. a norm $\|\cdot\|$ if for all x, y in the relative interior of the domain of f and $\alpha \in (0,1)$ we have*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \tfrac{1}{2}\beta\alpha(1-\alpha)\|x-y\|^2$$

We now define strong smoothness. Note that a strongly smooth function f is always finite.

Definition 2 *A function $f : \mathcal{X} \to \mathbb{R}$ is β-strongly smooth w.r.t. a norm $\|\cdot\|$ if f is everywhere differentiable and if for all x, y we have*

$$f(x+y) \leq f(x) + \langle \nabla f(x), y\rangle + \tfrac{1}{2}\beta\|y\|^2$$

The following theorem states that strong convexity and strong smoothness are dual properties. Recall that the biconjugate $f^{\star\star}$ equals f if and only if f is closed and convex.

Theorem 3 *(Strong/Smooth Duality) Assume that f is a closed and convex function. Then f is β-strongly convex w.r.t. a norm $\|\cdot\|$ if and only if $f^\star$ is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|\cdot\|_\star$.*

Subtly, note that while the domain of a strongly convex function f may be a proper subset of $\mathcal{X}$ (important for a number of settings), its conjugate $f^\star$ always has a domain which is $\mathcal{X}$ (since if $f^\star$ is strongly smooth then it is finite and everywhere differentiable). The above theorem can be found, for instance, in Zalinescu [2002] (see Corollary 3.5.11 on p. 217 and Remark 3.5.3 on p. 218). In the machine learning literature, a proof of one direction (strong convexity $\Rightarrow$ strong smoothness) can be found in Shalev-Shwartz [2007]. We could not find a proof of the reverse implication in a place easily accessible to machine learning people. So, a self-contained proof is provided in the appendix.

The following direct corollary of Theorem. 3 is central in proving both regret and generalization bounds.

Corollary 4 *If f is β strongly convex w.r.t. $\|\cdot\|$ and $f^\star(\mathbf{0}) = 0$, then, denoting the partial sum $\sum_{j\leq i} v_j$ by $v_{1:i}$, we have, for any sequence $v_1, \ldots, v_n$ and for any u,*

$$\sum_{i=1}^n \langle v_i, u\rangle - f(u) \leq f^\star(v_{1:n}) \leq \sum_{i=1}^n \langle \nabla f^\star(v_{1:i-1}), v_i\rangle + \frac{1}{2\beta}\sum_{i=1}^n \|v_i\|_\star^2 \,.$$

Proof: The 1st inequality is Fenchel-Young and the 2nd is from the definition of smoothness by induction. ■

2.3 Machine learning implications of the strong-convexity / strong-smoothness duality

We consider two learning models.

- **Online convex optimization:** Let $\mathcal{W}$ be a convex set. Online convex optimization is a two player repeated game. On round t of the game, the learner (first player) should choose $w_t \in \mathcal{W}$ and the environment (second player) responds with a convex function over $\mathcal{W}$, i.e. $l_t : \mathcal{W} \to \mathbb{R}$. The goal of the learner is to minimize its regret defined as:

$$\frac{1}{n}\sum_{t=1}^n l_t(w_t) - \min_{w\in\mathcal{W}} \frac{1}{n}\sum_{t=1}^n l_t(w) \,.$$

- **Batch learning of linear predictors:** Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Our goal is to learn a prediction rule from $\mathcal{X}$ to $\mathcal{Y}$. The prediction rule we use is based on a linear mapping $x \mapsto \langle w, x\rangle$, and the quality of the prediction is assessed by a loss function $l(\langle w, x\rangle, y)$. Our primary goal is to find w that has low risk (a.k.a. generalization error), defined as $L(w) = \mathbb{E}[l(\langle w, x\rangle, y)]$, where expectation is with respect to $\mathcal{D}$. To do so, we can sample n i.i.d. examples from $\mathcal{D}$ and observe the empirical risk, $\hat{L}(w) = \frac{1}{n}\sum_{i=1}^n l(\langle w, x_i\rangle, y_i)$. The goal of the learner is to find $\hat{w}$ with a low excess risk defined as:

$$L(\hat{w}) - \min_{w\in\mathcal{W}} L(w) \,,$$

  where $\mathcal{W}$ is a set of vectors that forms the comparison class.

We now seamlessly provide learning guarantees for both models based on Corollary 4. We start with the online convex optimization model.

Algorithm 1 Online Mirror Descent

$w_1 \leftarrow \nabla f^\star(\mathbf{0})$
for $t = 1$ to T **do**
  Play $w_t \in \mathcal{W}$
  Receive l_t and pick $v_t \in \partial l_t(w_t)$
  $w_{t+1} \leftarrow \nabla f^\star\left(-\eta \sum_{s=1}^t v_t\right)$
end for

Regret Bound for Online Convex Optimization Algorithm 1 provides one common algorithm which achieves the following regret bound. It is one of a family of algorithms that enjoy the same regret bound (see Shalev-Shwartz [2007]).

Theorem 5 *(Regret) Suppose Algorithm 1 is used with a function f that is β-strongly convex w.r.t. a norm $\|\cdot\|$ on $\mathcal{W}$ and has $f^\star(\mathbf{0}) = 0$. Suppose the loss functions l_t are convex and V-Lipschitz w.r.t. the dual norm $\|\cdot\|_\star$. Then, the algorithm run with any positive η enjoys the regret bound,*

$$\sum_{t=1}^T l_t(w_t) - \min_{u\in\mathcal{W}} \sum_{t=1}^T l_t(u) \leq \frac{\max_{u\in\mathcal{W}} f(u)}{\eta} + \frac{\eta V^2 T}{2\beta}$$

Proof: Apply Corollary 4 to the sequence $-\eta v_1, \ldots, -\eta v_T$ to get, for all u,

$$-\eta \sum_{t=1}^T \langle v_t, u\rangle - f(u) \leq -\eta \sum_{t=1}^T \langle v_t, w_t\rangle + \frac{1}{2\beta}\sum_{t=1}^T \|\eta v_t\|_\star^2 \,.$$

Using the fact that l_t is V-Lipschitz, we get $\|v_t\|_\star \leq V$. Plugging this into the inequality above and rearranging gives, $\sum_{t=1}^T \langle v_t, w_t - u\rangle \leq \frac{f(u)}{\eta} + \frac{\eta V^2 T}{2\beta}$. By convexity of l_t, $l_t(w_t) - l_t(u) \leq \langle v_t, w_t - u\rangle$. Therefore, $\sum_{t=1}^T l_t(w_t) - \sum_{t=1}^T l_t(u) \leq \frac{f(u)}{\eta} + \frac{\eta V^2 T}{2\beta}$. Since the above holds for all $u \in \mathcal{W}$ the result follows. ■

Generalization bound for the batch model via Rademacher analysis Let $\mathcal{T} = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X}\times\mathcal{Y})^n$ be a training set obtained by sampling i.i.d. examples from $\mathcal{D}$. For a class of real valued functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$, define its Rademacher complexity on $\mathcal{T}$ to be

$$\mathcal{R}_{\mathcal{T}}(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right] \,.$$

Here, the expectation is over ϵ_i's, which are i.i.d. Rademacher random variables, i.e. $\mathbb{P}(\epsilon_i = -1) = \mathbb{P}(\epsilon_1 = +1) = \frac{1}{2}$. It is well known that bounds on Rademacher complexity of a class immediately yield generalization bounds for classifiers picked from that class (assuming the loss function is Lipschitz). Recently, Kakade et al. [2008] proved Rademacher complexity bounds for classes consisting of linear predictors using strong convexity arguments. We now give a quick proof of their main result using Corollary 4. This proof is essentially the same as their original proof but highlights the importance of Corollary 4.

Theorem 6 *(Generalization) Let f be a β-strongly convex function w.r.t. a norm $\|\cdot\|$ and assume that $f^\star(\mathbf{0}) = 0$. Let $\mathcal{X} = \{x : \|x\|_\star \leq X\}$ and $\mathcal{W} = \{w : f(w) \leq f_{\max}\}$. Consider the class of linear functions, $\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathcal{W}\}$. Then, for any dataset $\mathcal{T} \in \mathcal{X}^n$, we have*

$$\mathcal{R}_{\mathcal{T}}(\mathcal{F}) \leq X\sqrt{\frac{2f_{\max}}{\beta n}} \,.$$

Proof: Let $\lambda > 0$. Apply Corollary 4 with $u = w$ and $v_i = \lambda\epsilon_i x_i$ to get,

$$\begin{aligned}\sup_{w\in\mathcal{W}} \sum_{i=1}^n \langle w, \lambda\epsilon_i x_i\rangle &\leq \frac{\lambda^2}{2\beta}\sum_{i=1}^n \|\epsilon_i x_i\|_\star^2 + \sup_{w\in\mathcal{W}} f(w) + \sum_{i=1}^n \langle \nabla f^\star(v_{1:i-1}), \epsilon_i x_i\rangle \\ &\leq \frac{\lambda^2 X^2 n}{2\beta} + f_{\max} + \sum_{i=1}^n \langle \nabla f^\star(v_{1:i-1}), \epsilon_i x_i\rangle \,.\end{aligned}$$

Now take expectation on both sides. The left hand side is $n\lambda\mathcal{R}_{\mathcal{T}}(\mathcal{F})$ and the last term on the right hand side becomes zero. Dividing throughout by $n\lambda$, we get, $\mathcal{R}_{\mathcal{T}}(\mathcal{F}) \leq \frac{\lambda X^2}{2\beta} + \frac{f_{\max}}{n\lambda}$. Optimizing over λ gives us the result. ■

Combining the above with the contraction lemma and standard Rademacher based generalization bounds (see e.g. Bartlett and Mendelson [2002], Kakade et al. [2008]) we obtain:

Corollary 7 *Let f be a β-strongly convex function w.r.t. a norm $\|\cdot\|$ and assume that $f^\star(\mathbf{0}) = 0$. Let $\mathcal{X} = \{x : \|x\|_\star \leq X\}$ and $\mathcal{W} = \{w : f(w) \leq f_{\max}\}$. Let l be an ρ-Lipschitz scalar loss function and let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Then, the algorithm that receives n i.i.d. examples and returns $\hat{w}$ that minimizes the empirical risk, $\hat{L}(w)$, satisfies*

$$\mathbb{E}\left[L(\hat{w}) - \min_{w\in\mathcal{W}} L(w)\right] \;\leq\; O\left(\rho\, X\sqrt{\frac{f_{\max}}{\beta n}}\right) ,$$

where expectation is with respect to the choice of the n i.i.d. examples.

We note that it is also easy to obtain a generalization bound that holds with high probability, but for simplicity of the presentation we stick to expectations.

2.4 Strongly Convex Matrix Functions

Before we consider strongly convex matrix functions, let us recall the following result about strong convexity of vector ℓ_p norm. Its proof can be found e.g. in Shalev-Shwartz [2007].

Lemma 8 *Let $q \in [1, 2]$. The function $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(w) = \frac{1}{2}\|w\|_q^2$ is $(q-1)$-strongly convex with respect to $\|\cdot\|_q$ over $\mathbb{R}^d$.*

We mainly use the above lemma to obtain results with respect to the norms $\|\cdot\|_2$ and $\|\cdot\|_1$. The case $q = 2$ is straightforward. Obtaining results with respect to $\|\cdot\|_1$ is slightly more tricky since for $q = 1$ the strong convexity parameter is 0 (meaning that the function is not strongly convex). To overcome this problem, we shall set q to be slightly more than 1, e.g. $q = \frac{\ln(d)}{\ln(d)-1}$. For this choice of q, the strong convexity parameter becomes $q - 1 = 1/(\ln(d) - 1) \geq 1/\ln(d)$ and the value of p corresponds to the dual norm is $p = (1 - 1/q)^{-1} = \ln(d)$. Note that for any $x \in \mathbb{R}^d$ we have

$$\|x\|_\infty \leq \|x\|_p \leq (d\|x\|_\infty^p)^{1/p} = d^{1/p}\|x\|_\infty = e\,\|x\|_\infty \leq 3\,\|x\|_\infty .$$

Hence the dual norms are also equivalent up to a factor of 3: $\|w\|_1 \geq \|w\|_q \geq \|w\|_1/3$. The above lemma therefore implies the following corollary.

Corollary 9 *The function $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(w) = \frac{1}{2}\|w\|_q^2$ for $q = \frac{\ln(d)}{\ln(d)-1}$ is $1/(3\ln(d))$-strongly convex with respect to $\|\cdot\|_1$ over $\mathbb{R}^d$.*

We now consider two families of strongly convex matrix functions.

Schatten q-norms The first result we need is the counterpart of Lemma. 8 for the q-Schatten norm defined as $\|\mathbf{X}\|_{S(q)} := \|\sigma(\mathbf{X})\|_q$ This result can be found in Ball et al. [1994].

Theorem 10 *(Schatten matrix functions) Let $q \in [1, 2]$. The function $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ defined as $F(\mathbf{X}) = \frac{1}{2}\|\sigma(\mathbf{X})\|_q^2$ is $(q-1)$-strongly convex w.r.t. the q-Schatten norm $\|\mathbf{X}\|_{S(q)} := \|\sigma(\mathbf{X})\|_q$ over $\mathbb{R}^{m\times n}$.*

As above, choosing q to be $\frac{\ln m'}{\ln(m')-1}$ for $m' = \min\{m, n\}$ gives the following corollary.

Corollary 11 *The function $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ defined as $F(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_{S(q)}^2$ for $q = \frac{\ln(m')}{\ln(m')-1}$ is $1/(3\ln(m'))$-strongly convex with respect to $\|\cdot\|_{S(1)}$ over $\mathbb{R}^{m\times n}$.*

Group Norms. Let $\mathbf{X} = (\mathbf{X}^1 \mathbf{X}^2 \dots \mathbf{X}^n)$ be a $m \times n$ real matrix with *columns* $\mathbf{X}^i \in \mathbb{R}^m$. We denote by $\|\mathbf{X}\|_{r,p}$ as

$$\|\mathbf{X}\|_{r,p} := \| \left(\|\mathbf{X}^1\|_r, \dots, \|\mathbf{X}^n\|_r\right) \|_p \, .$$

That is, we apply $\|\cdot\|_r$ to each column of $\mathbf{X}$ to get a vector in $\mathbb{R}^n$ to which we apply the norm $\|\cdot\|_p$ to get the value of $\|\mathbf{X}\|_{r,p}$. It is easy to check that this is indeed a norm. The dual of $\|\cdot\|_{r,p}$ is $\|\cdot\|_{s,t}$ where $1/r+1/s = 1$ and $1/p+1/t = 1$. The following theorem, which appears in a slightly weaker form in Juditsky and Nemirovski [2008], provides us with an easy way to construct strongly convex group norms. We provide a proof in the appendix which is much simpler than that of Juditsky and Nemirovski [2008] and is completely "calculus free".

Theorem 12 *(Group Norms) Let Ψ, Φ be absolutely symmetric norms on $\mathbb{R}^m, \mathbb{R}^n$. Let $\Phi^2 \circ \sqrt{} : \mathbb{R}^n \to \mathbb{R}^*$ denote the following function,*

$$(\Phi^2 \circ \sqrt{})(x) := \Phi^2(\sqrt{x_1}, \dots, \sqrt{x_n}) \, . \tag{1}$$

Suppose, $(\Phi^2 \circ \sqrt{})$ is a norm on $\mathbb{R}^n$. Further, let the functions Ψ^2 and Φ^2 be σ_1- and σ_2-smooth w.r.t. Ψ and Φ respectively. Then, $\|\cdot\|^2_{\Psi,\Phi}$ is $(\sigma_1 + \sigma_2)$-smooth w.r.t. $\|\cdot\|_{\Psi,\Phi}$.

The condition that Eq. (1) be a norm appears strange but in fact it already occurs in the literature. Norms satisfying it are called *quadratic symmetric gauge* functions (or Q-norms) [Bhatia, 1997, p. 89]. It is easy to see that $\|\cdot\|_p$ for $p \geq 2$ is a Q-norm. Now using strong convexity/strong smoothness duality and the discussion preceding Corollary 9, we get the following corollary.

Corollary 13 *The function $F : \mathbb{R}^{m \times n} \to \mathbb{R}$ defined as $F(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|^2_{2,q}$ for $q = \frac{\ln(n)}{\ln(n)-1}$ is $1/(3\ln(n))$-strongly convex with respect to $\|\cdot\|_{2,1}$ over $\mathbb{R}^{m \times n}$.*

2.5 Putting it all together

Combining Lemma. 8 and Corollary 9 with the bounds given in Theorem. 5 and Corollary 7 we therefore obtain the following two corollaries.

Corollary 14 *Let $\mathcal{W} = \{w : \|w\|_1 \leq W\}$ and let $l_1, \dots, l_n$ be a sequence of functions which are X-Lipschitz w.r.t. $\|\cdot\|_\infty$. Then, there exists an online algorithm with a regret bound of the form*

$$\frac{1}{n}\sum_{t=1}^n l_t(w_t) - \min_{w \in \mathcal{W}} \frac{1}{n}\sum_{t=1}^n l_t(w) \;\leq\; O\left(X\,W\sqrt{\frac{\ln(d)}{n}}\right) \, .$$

Corollary 15 *Let $\mathcal{W} = \{w : \|w\|_1 \leq W\}$ and let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq X\}$. Let l be an ρ-Lipschitz scalar loss function and let $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$. Then, there exists a batch learning algorithm that returns a vector $\hat{w}$ such that*

$$\mathbb{E}\left[L(\hat{w}) - \min_{w \in \mathcal{W}} L(w)\right] \;\leq\; O\left(X\,W\sqrt{\frac{\ln(d)}{n}}\right) \, .$$

Results of the same flavor can be obtained for learning matrices. For simplicity, we present the following two corollaries only for the online model, but it is easy to derive their batch counterparts.

Corollary 16 *Let $\mathcal{W} = \{\mathbf{W} \in \mathbb{R}^{k \times d} : \|\mathbf{W}\|_{2,1} \leq W\}$ and let $l_1, \dots, l_n$ be a sequence of functions which are X-Lipschitz w.r.t. $\|\cdot\|_{2,\infty}$. Then, there exists an online algorithm with a regret bound of the form*

$$\frac{1}{n}\sum_{t=1}^n l_t(\mathbf{W}_t) - \min_{\mathbf{W} \in \mathcal{W}} \frac{1}{n}\sum_{t=1}^n l_t(\mathbf{W}) \;\leq\; O\left(X\,W\sqrt{\frac{\ln(d)}{n}}\right) \, .$$

Corollary 17 *Let $\mathcal{W} = \{\mathbf{W} \in \mathbb{R}^{k \times d} : \|\mathbf{W}\|_{S(1)} \leq W\}$ and let $l_1, \dots, l_n$ be a sequence of functions which are X-Lipschitz w.r.t. $\|\cdot\|_{S(\infty)}$. Then, there exists an online algorithm with a regret bound of the form*

$$\frac{1}{n}\sum_{t=1}^n l_t(\mathbf{W}_t) - \min_{\mathbf{W} \in \mathcal{W}} \frac{1}{n}\sum_{t=1}^n l_t(\mathbf{W}) \;\leq\; O\left(X\,W\sqrt{\frac{\ln(\min\{k,d\})}{n}}\right) \, .$$

3 Matrix Regularization

We are now ready to demonstrate the power of the general techniques we derived in the previous section. Consider a learning problem (either online or batch) in which $\mathcal{X}$ is a subset of a matrix space (of dimension $k \times d$) and we would like to learn a linear predictor of the form $\mathbf{X} \mapsto \langle \mathbf{W}, \mathbf{X} \rangle$ where $\mathbf{W}$ is also a matrix of the same dimension. The loss function takes the form $l(\langle \mathbf{W}, \mathbf{X} \rangle, y)$ and we assume for simplicity that l is 1-Lipschitz with respect to its first argument. For example, l can be the absolute loss, $l(a, y) = |a - y|$, or the hinge-loss, $l(a, y) = \max\{0, 1 - ya\}$.

For the sake of concreteness, let us focus on the batch learning setting, but we note that the discussion below is relevant to the online learning model as well. Our prior knowledge on the learning problem is encoded by the definition of the comparison class $\mathcal{W}$ that we use. In particular, all the comparison classes we use take the form $\mathcal{W} = \{\mathbf{W} : \|\mathbf{W}\| \leq W\}$, where the only difference is what norm do we use. We shall compare the following four classes:

$$\mathcal{W}_{1,1} = \{\mathbf{W} : \|\mathbf{W}\|_{1,1} \leq W_{1,1}\} \qquad \mathcal{W}_{2,2} = \{\mathbf{W} : \|\mathbf{W}\|_{2,2} \leq W_{2,2}\}$$
$$\mathcal{W}_{2,1} = \{\mathbf{W} : \|\mathbf{W}\|_{2,1} \leq W_{2,1}\} \qquad \mathcal{W}_{S(1)} = \{\mathbf{W} : \|\mathbf{W}\|_{S(1)} \leq W_{S(1)}\}$$

Let us denote $X_{\infty,\infty} = \sup_{x \in \mathcal{X}} \|\mathbf{X}\|_{\infty,\infty}$. We define $X_{2,2}, X_{2,\infty}, X_{S(\infty)}$ analogously. Applying the results of the previous section to these classes we obtain the bounds given in Table 1 where for simplicity we ignore constants.

class	$\mathcal{W}_{1,1}$	$\mathcal{W}_{2,2}$	$\mathcal{W}_{2,1}$	$\mathcal{W}_{S(1)}$
bound	$W_{1,1}\, X_{\infty,\infty} \sqrt{\frac{\ln(kd)}{n}}$	$W_{2,2}\, X_{2,2} \sqrt{\frac{1}{n}}$	$W_{2,1}\, X_{2,\infty} \sqrt{\frac{\ln(d)}{n}}$	$W_{S(1)}\, X_{S(\infty)} \sqrt{\frac{\ln(\min\{d,k\})}{n}}$

Table 1: List of bounds for learning with matrices. For simplicity we ignore constants.

Let us now discuss which class should be used based on prior knowledge on properties of the learning problem. We start with the well known difference between $\mathcal{W}_{1,1}$ and $\mathcal{W}_{2,2}$. Note that both of these classes ignore the fact that $\mathbf{W}$ is organized as a $k \times d$ matrix and simply refer to $\mathbf{W}$ as a single vector of dimension kd. The difference between $\mathcal{W}_{1,1}$ and $\mathcal{W}_{2,2}$ is therefore the usual difference between ℓ_1 and ℓ_2 regularization. To understand this difference, suppose that $\mathbf{W}$ is some matrix that performs well on the distribution we have. Then, we should take the radius of each class to be the minimal possible while still containing $\mathbf{W}$, namely, either $\|\mathbf{W}\|_{1,1}$ or $\|\mathbf{W}\|_{2,2}$. Clearly, $\|\mathbf{W}\|_{2,2} \leq \|\mathbf{W}\|_{1,1}$ and therefore in terms of this term there is a clear advantage to use the class $\mathcal{W}_{2,2}$. On the other hand, $X_{2,2} \geq X_{\infty,\infty}$. We therefore need to understand which of these inequalities is more important. Of course, in general, the answer to this question is data dependent. However, we can isolate properties of the distribution that can help us choose the better class.

One useful property is sparsity of either $\mathbf{X}$ or $\mathbf{W}$. If $\mathbf{X}$ is assumed to be s sparse (i.e., it has at most s non-zero elements), then we have $X_{2,2} \leq \sqrt{s} X_{\infty,\infty}$. That is, for a small s, the difference between $X_{2,2}$ and $X_{\infty,\infty}$ is small. In contrast, if $\mathbf{X}$ is very dense and each of its entries is bounded away from zero, e.g. $\mathcal{X} \in \{\pm 1\}^{k \times d}$, then $\|\mathbf{X}\|_{2,2} = \sqrt{kd} \|\mathbf{X}\|_{\infty,\infty}$. The same arguments are true for $\mathbf{W}$. Hence, with prior knowledge about the sparsity of $\mathbf{X}$ and $\mathbf{W}$ we can guess which of the bounds will be smaller.

Next, we tackle the more interesting cases of $\mathcal{W}_{2,1}$ and $\mathcal{W}_{S(1)}$. For the former, recall that we first apply ℓ_2 norm on each column of $\mathbf{W}$ and then apply ℓ_1 norm on the obtained vector of norm values. Similarly, to calculate $\|\mathbf{X}\|_{2,\infty}$ we first apply ℓ_2 norm on columns of $\mathbf{X}$ and then apply ℓ_∞ norm on the obtained vector of norm values. Let us now compare $\mathcal{W}_{2,1}$ to $\mathcal{W}_{1,1}$. Suppose that the columns of $\mathbf{X}$ are very sparse. Therefore, the ℓ_2 norm of each column of $\mathbf{X}$ is very close to its ℓ_∞ norm. On the other hand, if some of the columns of $\mathbf{W}$ are dense, then $\|\mathbf{W}\|_{2,1}$ can be order of $\sqrt{k}$ smaller than $\|\mathbf{W}\|_{1,1}$. In that case, the class $\mathcal{W}_{2,1}$ is preferable over the class $\mathcal{W}_{1,1}$. As we show later, this is the case in multi-class problems, and we shall indeed present an improved multi-class algorithm that uses the class $\mathcal{W}_{2,1}$. Of course, in some problems, columns of $\mathbf{X}$ might be very dense while columns of $\mathbf{W}$ can be sparse. In such cases, using $\mathcal{W}_{1,1}$ is better than using $\mathcal{W}_{2,1}$.

Now lets compare $\mathcal{W}_{2,1}$ to $\mathcal{W}_{2,2}$. Similarly to the previous discussion, choosing $\mathcal{W}_{2,1}$ over $\mathcal{W}_{2,2}$ makes sense if we assume that the vector of ℓ_2 norms of columns, $(\|\mathbf{W}^1\|_2, \ldots, \|\mathbf{W}^d\|_2)$, is sparse. This implies that we assume a "group"-sparsity pattern of $\mathbf{W}$, i.e., each column of $\mathbf{W}$ is either the all zeros column or is dense. This type of grouped-sparsity has been studied in the context of group Lasso and multi-task learning. Indeed, we present bounds for multi-task learning that relies on this assumption. Without the group-sparsity assumption, it might be better to use $\mathcal{W}_{2,2}$ over $\mathcal{W}_{2,1}$.

Finally, we discuss when it makes sense to use $\mathcal{W}_{S(1)}$. Recall that $\|\mathbf{W}\|_{S(1)} = \|\sigma(\mathbf{W})\|_1$, where $\sigma(\mathbf{W})$ is the vector of singular values of $\mathbf{W}$, and $\|\mathbf{X}\|_{S(\infty)} = \|\sigma(\mathbf{X})\|_\infty$. Therefore, the class $\mathcal{W}_{S(1)}$ should be

used when we assume that the *spectrum* of $\mathbf{W}$ is sparse while the spectrum of $\mathbf{X}$ is dense. This means that the prior knowledge we employ is that $\mathbf{W}$ is of low rank while $\mathcal{X}$ is of high rank. Note that $\mathcal{W}_{2,2}$ can be defined equivalently as $\mathcal{W}_{S(2)}$. Therefore, the difference between $\mathcal{W}_{S(1)}$ and $\mathcal{W}_{2,2}$ is similar to the difference between $\mathcal{W}_{1,1}$ and $\mathcal{W}_{2,2}$ just that instead of considering sparsity properties of the elements of $\mathbf{W}$ and $\mathbf{X}$ we consider sparsity properties of the spectrum of $\mathbf{W}$ and $\mathbf{X}$.

In the next sections we demonstrate how to apply the general methodology described above in order to derive a few generalization and regret bounds for problems of recent interest.

4 Multi-task learning

Suppose we are simultaneously solving k-multivariate prediction problems, where each learning example is of the form $(\mathbf{X}, \mathbf{y})$ where $\mathbf{X} \in \mathbb{R}^{k \times d}$ is a matrix of example vectors with examples from different tasks sitting in rows of $\mathbf{X}$, and $\mathbf{y} \in \mathbb{R}^k$ are the responses for the k problems. To predict the k responses, we learn a matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$ such that $\mathrm{Diag}(\mathbf{W}^\top \mathbf{X})$ is a good predictor of $\mathbf{y}$. In this section, we denote *row* j of $\mathbf{W}$ by $\mathbf{w}^j$. The predictor for the jth task is therefore $\mathbf{w}^j$. The quality of a prediction $\left\langle \mathbf{w}^j, \mathbf{x}^j \right\rangle$ for the j'th task is assessed by a loss function $l^j : \mathbb{R} \times \mathcal{Y}^j \to \mathbb{R}$; And, the total loss of $\mathbf{W}$ on an example $(\mathbf{X}, \mathbf{y})$ is defined to be the sum of the individual losses,

$$l(\mathbf{W}, \mathbf{X}, \mathbf{y}) = \sum_{j=1}^{k} l^j(\left\langle \mathbf{w}^j, \mathbf{x}^j \right\rangle, y^j) \ .$$

This formulation allows us to mix regression and classification problems and even use different loss functions for different tasks. Such "heterogeneous" multi-task learning has attracted recent attention [Yang et al., 2009].

If the tasks are related, then it is natural to use regularizers that "couple" the tasks together so that similarities across tasks can be exploited. Considerations of common sparsity patterns (same features relevant across different tasks) lead to the use of group norm regularizers (i.e. using the comparison class $\mathcal{W}_{2,1}$ defined in the previous section) while rank considerations (the $\mathbf{w}^j$'s lie in a low dimensional linear space) lead to the use of unitarily invariant norms as regularizers (i.e. the comparison class is $\mathcal{W}_{S(1)}$).

We now describe online and batch multi-task learning using different matrix norm.

4.1 Online multi-task learning

In the online model, on round t the learner first uses $\mathbf{W}_t$ to predict the vector of responses and then it pays the cost $l_t(\mathbf{W}_t) = l(\mathbf{W}_t, \mathbf{X}_t, \mathbf{y}_t) = \sum_{j=1}^{k} l^j\left(\left\langle \mathbf{w}_t^j, \mathbf{x}_t^j \right\rangle, y_t^j\right)$. Let $\mathbf{V}_t \in \mathbb{R}^{k \times d}$ be a sub-gradient of l_t at $\mathbf{W}_t$. It is easy to verify that the j'th row of $\mathbf{V}_t$, denoted $\mathbf{v}_t^j$, is a sub-gradient of $l^j\left(\left\langle \mathbf{w}_t^j, \mathbf{x}_t^j \right\rangle, y_t^j\right)$ at $\mathbf{w}_t^j$. Assuming that l^j is ρ-Lipschitz with respect to its first argument, we obtain that $\mathbf{v}_t^j = \tau_t^j \mathbf{x}_t^j$ for some $\tau_t^j \in [-\rho, \rho]$. In other words, $\mathbf{V}_t = \mathrm{Diag}(\tau_t)\, \mathbf{X}_t$. It is easy to verify that $\|\mathbf{V}_t\|_{r,p} \leq \rho \, \|\mathbf{X}\|_{r,p}$ for any $r, p \geq 1$. In addition, since any Schatten norm is sub-multiplicative we also have that $\|\mathbf{V}_t\|_{S(\infty)} \leq \|\mathrm{Diag}(\tau_t)\|_{S(\infty)} \, \|\mathbf{X}_t\|_{S(\infty)} \leq \rho \, \|\mathbf{X}_t\|_{S(\infty)}$. We therefore obtain the following:

Corollary 18 *Let $\mathcal{W}_{1,1}, \mathcal{W}_{2,2}, \mathcal{W}_{2,1}, \mathcal{W}_{S(1)}$ be the classes defined in Section 3 and lets $X_{\infty,\infty}, X_{2,2}, X_{2,\infty}, X_{S(\infty)}$ be the radius of $\mathcal{X}$ w.r.t. the corresponding norms. Then, there exist online multi-task learning algorithms with regret bounds according to Table 1.*

Let us now discuss few implications of these bounds, and for simplicity assume that $k < d$. Recall that each column of $\mathbf{X}$ represents the value of a single feature for all the tasks. As discussed in the previous section, if the matrix $\mathbf{X}$ is dense and if we assume that $\mathbf{W}$ is sparse, then using the class $\mathcal{W}_{1,1}$ is better than using $\mathcal{W}_{2,2}$. Such a scenario often happens when we have many irrelevant features and only are few features that can predict the target reasonably well. Concretely, suppose that $\mathbf{X} \in \{0,1\}^{k \times d}$ and that it typically has s_x non-zero values. Suppose also that there exists a matrix $\mathbf{W}$ that predicts the targets of the different tasks reasonably well and has s_w non-zero values. Then, the bound for $\mathcal{W}_{1,1}$ is order of $s_w \sqrt{\ln(dk)/n}$ while the bound for $\mathcal{W}_{2,2}$ is order of $\sqrt{s_w \, s_x / n}$. Thus, $\mathcal{W}_{1,1}$ will be better if $s_w < s_x / \ln(dk)$.

Now, consider the class $\mathcal{W}_{2,1}$. Let us further assume the following. The non-zero elements of $\mathbf{W}$ are grouped into s_g columns and are roughly distributed evenly over those columns; The non-zeros of $\mathbf{X}$ are roughly distributed evenly over the columns. Then, the bound for $\mathcal{W}_{2,1}$ is $s_g \sqrt{(s_w/s_g)\,(s_x/d)\,\ln(d)/n} = \sqrt{s_g \, s_w \, (s_x/d) \, \ln(d)/n}$. This bound will be better than the bound of $\mathcal{W}_{2,2}$ if $s_g \ln(d) < d$ and will be better than the bound of $\mathcal{W}_{1,1}$ if $s_g s_x / d < s_w$. We see that there are scenarios in which the group norm is better than the non-grouped norms and that the most adequate class depends on properties of the problem and our prior beliefs on a good predictor $\mathbf{W}$.

As to the bound for $\mathcal{W}_{S(1)}$, it is easy to verify that if the rows of $\mathbf{W}$ sits in a low dimensional subspace then the spectrum of $\mathbf{W}$ will be sparse. Similarly, the value of $\|\mathbf{X}\|_{S(\infty)}$ depends on the maximal singular value of $\mathbf{X}$, which is likely to be small if we assume that all the "energy" of $\mathbf{X}$ is spread over its entire spectrum. In such cases, $\mathcal{W}_{S(1)}$ can be the best choice. This is an example of a different type of prior knowledge on the problem.

4.2 Batch multi-task learning

In the batch setting we see a dataset $\mathcal{T} = ((\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_n, \mathbf{y}_n))$ consisting of i.i.d. samples drawn from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. In the k-task setting, $\mathcal{X} \subseteq \mathbb{R}^{k \times d}$. Analogous to the single task case, we define the risk and empirical risk of a multitask predictor $\mathbf{W} \in \mathbb{R}^{k \times d}$ as:

$$\widehat{L}(\mathbf{W}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} l^j \left(\left\langle \mathbf{w}^j, \mathbf{X}_i^j \right\rangle, y_i^j \right) \quad ; \quad L(\mathbf{W}) := \mathbb{E}_{(\mathbf{X},\mathbf{y}) \sim \mathcal{D}} \left[\sum_{j=1}^{k} l^j \left(\langle \mathbf{w}^j, \mathbf{X}^j \rangle, y^j \right) \right] .$$

Let $\mathcal{W}$ be some class of matrices, and define the empirical risk minimizer, $\widehat{\mathbf{W}} := \operatorname{argmin}_{\mathbf{W} \in \mathcal{W}} \widehat{L}(\widehat{\mathbf{W}})$. To obtain excess risk bounds for $\widehat{\mathbf{W}}$, we need to consider the k-task Rademacher complexity

$$\mathcal{R}_{\mathcal{T}}^k(\mathcal{W}) := \mathbb{E} \left[\sup_{\mathbf{W} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \epsilon_i^j \left\langle \mathbf{w}^j, \mathbf{X}_i^j \right\rangle \right] .$$

because, assuming each l^j is ρ-Lipschitz, we have the bound $\mathbb{E}\left[L(\widehat{\mathbf{W}}) - \min_{\mathbf{W} \in \mathcal{W}} L(\mathbf{W}) \right] \leq \rho \mathbb{E}\left[\mathcal{R}_{\mathcal{T}}^k(\mathcal{W}) \right]$. This bound follows easily from Talagrand's contraction inequality and Thm. 8 in Maurer [2006]. We can use matrix strong convexity to give the following k-task Rademacher bound.

Theorem 19 *(Multitask Generalization) Suppose $F(\mathbf{W}) \leq f_{\max}$ for all $\mathbf{W} \in \mathcal{W}$ for a function F that is β-strongly convex w.r.t. some (matrix) norm $\|\cdot\|$. If the norm $\|\cdot\|_\star$ is invariant under sign changes of the rows of its argument matrix then, for any dataset $\mathcal{T}$, we have, $\mathcal{R}_{\mathcal{T}}^k(\mathcal{W}) \leq X \sqrt{\frac{2 f_{\max}}{\beta n}}$, where X is an upper bound on $\|\mathbf{X}_i\|_\star$.*

Proof: We can rewrite $\mathcal{R}_{\mathcal{T}}^k(\mathcal{W})$ as

$$\mathbb{E}\left[\sup_{\mathbf{W} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \epsilon_i^j \left\langle \mathbf{w}^j, \mathbf{X}_i^j \right\rangle \right] = \mathbb{E}\left[\sup_{\mathbf{W} \in \mathcal{W}} \frac{1}{n} \sum_{j=1}^{k} \left\langle \mathbf{w}^j, \sum_{i=1}^{n} \epsilon_i^j \mathbf{X}_i^j \right\rangle \right] = \mathbb{E}\left[\sup_{\mathbf{W} \in \mathcal{W}} \frac{1}{n} \left\langle \mathbf{W}, \sum_{i=1}^{n} \tilde{\mathbf{X}}_i \right\rangle \right] ,$$

where $\tilde{\mathbf{X}}_i \in \mathbb{R}^{k \times d}$ is defined by $\tilde{\mathbf{X}}_i^j = \epsilon_i^j \mathbf{X}_i^j$ and we have switched to a matrix inner product in the last line. By the assumption on the dual norm $\|\cdot\|_\star$, $\|\tilde{\mathbf{X}}_i\|_\star = \|\mathbf{X}_i\|_\star \leq X$. Now using Corollary 4 and proceeding as in the proof of Theorem. 6, we get, for any $\lambda > 0$, $\mathcal{R}_{\mathcal{T}}^k(\mathcal{W}) \leq \left(\frac{f_{\max}}{\lambda n} + \frac{\lambda X^2}{2\beta} \right)$. Optimizing over λ proves the theorem. ■

Note that both group (r,p)-norms and Schatten-p norms satisfy the invariance under row flips mentioned in the theorem above. Thus, we get the following corollary.

Corollary 20 *Let $\mathcal{W}_{1,1}, \mathcal{W}_{2,2}, \mathcal{W}_{2,1}, \mathcal{W}_{S(1)}$ be the classes defined in Section 3 and lets $X_{\infty,\infty}, X_{2,2}, X_{2,\infty}, X_{S(\infty)}$ be the radius of $\mathcal{X}$ w.r.t. the corresponding norms. Then, the (expected) excess multitask risk of the empirical multitask risk minimizer $\widehat{\mathbf{W}}$ satisfies the same bounds given in Table 1.*

5 Multi-class learning

In this section we consider multi-class categorization problems. We focus on the online learning model. On round t, the online algorithm receives an instance $x_t \in \mathbb{R}^d$ and is required to predict its label as a number in $\{1, \ldots, k\}$. Following the construction of Crammer and Singer [2000], the prediction is based on a matrix $\mathbf{W}_t \in \mathbb{R}^{k \times d}$ and is defined as the index of the maximal element of the vector $\mathbf{W}_t x_t$. We use the hinge-loss function adapted to the multi-class setting. That is,

$$l_t(\mathbf{W}_t) \;=\; \max_r (\mathbf{1}_{[r \neq y_t]} - (\langle \mathbf{w}_t^{y_t}, x_t \rangle - \langle \mathbf{w}_t^r, x_t \rangle)) = \max_r (\mathbf{1}_{[r \neq y_t]} - (\langle \mathbf{W}, \mathbf{X}_t^{r, y_t} \rangle)) ,$$

where $\mathbf{X}_t^{r,y_t}$ is a matrix with x_t on the y'th row, $-x_t$ on the r'th row, and zeros in all other elements. It is easy to verify that $l_t(\mathbf{W}_t)$ upper bounds the zero-one loss, i.e. if the prediction of $\mathbf{W}_t$ is r then $l_t(\mathbf{W}_t) \geq \mathbf{1}_{[r \neq y_t]}$.

A sub-gradient of $l_t(\mathbf{W}_t)$ is either a matrix of the form $-\mathbf{X}_t^{r,y_t}$ or the all zeros matrix. Note that each column of $\mathbf{X}_t^{r,y_t}$ is very sparse (contains only two elements). Therefore,

$$\|\mathbf{X}_t^{r,y_t}\|_{\infty,\infty} = \|x_t\|_\infty \;;\; \|\mathbf{X}_t^{r,y_t}\|_{2,2} = \sqrt{2}\,\|x_t\|_2 \;;\; \|\mathbf{X}_t^{r,y_t}\|_{2,\infty} = \sqrt{2}\,\|x_t\|_\infty \;;\; \|\mathbf{X}_t^{r,y_t}\|_{S(\infty)} = \sqrt{2}\,\|x_t\|_2$$

Based on this fact, we can easily obtain the following.

Corollary 21 *Let $\mathcal{W}_{1,1}, \mathcal{W}_{2,2}, \mathcal{W}_{2,1}, \mathcal{W}_{S(1)}$ be the classes defined in Section 3 and let $X_2 = \max_t \|x_t\|_2$ and $X_\infty = \max_t \|x_t\|_\infty$. Then, there exist online multi-class learning algorithms with regret bounds given by the following table*

class	$\mathcal{W}_{1,1}$	$\mathcal{W}_{2,2}$	$\mathcal{W}_{2,1}$	$\mathcal{W}_{S(1)}$
bound	$W_{1,1}\, X_\infty \sqrt{\frac{\ln(kd)}{n}}$	$W_{2,2}\, X_2 \sqrt{\frac{1}{n}}$	$W_{2,1}\, X_\infty \sqrt{\frac{\ln(d)}{n}}$	$W_{S(1)}\, X_2 \sqrt{\frac{\ln(\min\{d,k\})}{n}}$

Let us now discuss the implications of this bound. First, if $X_2 \approx X_\infty$, which will happen if instance vectors are sparse, then $\mathcal{W}_{1,1}$ and $\mathcal{W}_{2,1}$ will be inferior to $\mathcal{W}_{2,2}$. In such a case, using $\mathcal{W}_{S(1)}$ can be even better if $\mathbf{W}$ sits in a low dimensional space but each row of $\mathbf{W}$ still has a unit norm. Using $\mathcal{W}_{S(1)}$ in such a case was previously suggested by Amit et al. [2007], who observed that empirically, the class $\mathcal{W}_{S(1)}$ performs better than $\mathcal{W}_{2,2}$ when there is a shared structure between classes. The analysis given in Corollary 21 provides a first rigorous explanation to such a behavior.

Second, if X_2 is much larger than X_∞, and if columns of $\mathbf{W}$ share common sparsity pattern, then $\mathcal{W}_{2,1}$ can be factor of $\sqrt{k}$ better than $\mathcal{W}_{1,1}$ and factor of $\sqrt{d}$ better than $\mathcal{W}_{2,2}$. To demonstrate this, let us assume that each vector x_t is in $\{\pm 1\}^d$ and it represents experts advice of d experts. Therefore, $X_2 = \sqrt{d}\, X_\infty$. Next, assume that a combination of the advice of $s \ll d$ experts predicts very well the correct label (e.g., the label is represented by the binary number obtained from the advice of $s = \log(k)$ experts). In that case, W will be a matrix such that all of its columns will be 0 except s columns which will take values in $\{\pm 1\}$. The bounds for $\mathcal{W}_{1,1}, \mathcal{W}_{2,2}$, and $\mathcal{W}_{2,1}$ in that case becomes $ks\sqrt{\ln(kd)}$, $\sqrt{ksd}$, and $\sqrt{ks\ln(d)}$ respectively. That is, $\mathcal{W}_{2,1}$ is a factor of $\sqrt{ks}$ better than $\mathcal{W}_{1,1}$ and a factor of $\sqrt{d}$ better than $\mathcal{W}_{2,2}$ (ignoring logarithmic terms). The class $\mathcal{W}_{S(1)}$ will also have a dependent on $\sqrt{d}$ in such a case and thus it will be much worse than $\mathcal{W}_{2,2}$ when d is large.

For concreteness, we now utilize our result for deriving a group Multi-class Perceptron algorithm. To the best of our knowledge, this algorithm is new, and based on the discussion above, it should outperform both the multi-class Perceptron of Crammer and Singer [2000] as well as the vanilla application of the p-norm Perceptron framework of Gentile [2003], Grove et al. [2001] for multi-class categorization.

The algorithm is a specification of the general online mirror descent procedure (Algorithm 1) with $f(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_{2,r}^2$, $r = \log(d)/(\log(d)-1)$, and with a conservative update (i.e., we ignore rounds on which no prediction mistake has been made). Recall that the Fenchel dual function is $f^\star(\mathbf{V}) = \frac{1}{2}\|\mathbf{V}\|_{2,p}^2$ where $p = (1-1/r)^{-1} = \log(d)$. The (i,j) element of the gradient of $f^\star$ is

$$(\nabla f^\star(\mathbf{V}))_{i,j} \;=\; \frac{\|\mathbf{V}^j\|_2^{p-2}}{\|\mathbf{V}\|_{2,p}^{p-2}}\, V_{i,j}\;. \tag{2}$$

Algorithm 2 Group Multi-class Perceptron

$p = \log d$
$\mathbf{V}_1 = \mathbf{0} \in \mathbb{R}^{k\times d}$
for $t = 1, \ldots, T$ **do**
  Set $\mathbf{W}_t = \nabla f^\star(\mathbf{V}_t)$ (as defined in Eq. (2))
  Receive $\mathbf{x}_t \in \mathbb{R}^d$
  $\hat{y}_t = \arg\max_{r\in[k]} \left(\mathbf{W}_t \mathbf{x}_t\right)_r$
  Predict $\hat{y}_t$ and receive true label y_t
  $\mathbf{U}_t \in \mathbb{R}^{k\times d}$ is the matrix with $\mathbf{x}_t$ in the $\hat{y}_t$ row and $-\mathbf{x}_t$ in the y_t row
  Update: $\mathbf{V}_{t+1} = \mathbf{V}_t - \mathbf{U}_t$
end for

To analyze the performance of Algorithm 2, let $I \subseteq [n]$ be the set of rounds on which the algorithm made a prediction mistake. Note that the above algorithm is equivalent (in terms of the number of mistakes) to an algorithm that performs the update $\mathbf{V}_{t+1} = \mathbf{V}_t + \eta \mathbf{U}_t$ for any η (see Gentile [2003]). Therefore, we can

apply our general online regret bound (Corollary 16) on the sequence of examples in I we obtain that for any $\mathbf{W}$

$$\sum_{t\in I} l_t(\mathbf{W}_t) - \sum_{t\in I} l_t(\mathbf{W}) \;\le\; O\left(X_\infty \|\mathbf{W}\|_{2,1} \sqrt{\log(d)\,|I|}\right) \;.$$

Recall that $l_t(\mathbf{W}_t)$ upper bounds the zero-one error and therefore the above implies that

$$|I| - \sum_{t\in I} l_t(\mathbf{W}) \;\le\; O\left(X_\infty \|\mathbf{W}\|_{2,1} \sqrt{\log(d)\,|I|}\right) \;.$$

Solving for $|I|$ we conclude that:

Corollary 22 *The number of mistakes Algorithm 2 will make on any sequence of examples for which $\|x_t\|_\infty \le X_\infty$ is upper bounded by*

$$\min_{\mathbf{W}} \;\; \sum_t l_t(\mathbf{W}) + O\left(X_\infty \|\mathbf{W}\|_{2,1} \sqrt{\log(d) \sum_t L_t(\mathbf{W})}\right) \;\; .$$

6 Kernel learning

We briefly review the kernel learning setting first explored in Lanckriet et al. [2004]. Let $\mathcal{X}$ be an input space and let $\mathcal{T} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathcal{X}^n$ be the training dataset. Kernel algorithms work with the space of linear functions, $\{\mathbf{x} \mapsto \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) \;:\; \alpha_i \in \mathbb{R}\}$. In kernel learning, we consider a kernel *family* $\mathcal{K}$ and consider the class, $\{\mathbf{x} \mapsto \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \mathbf{x}) \;:\; K \in \mathcal{K},\; \alpha_i \in \mathbb{R}\}$. In particular, we can choose a finite set $\{K_1, \ldots, K_k\}$ of base kernels and consider the convex combinations, $\mathcal{K}_c^+ = \left\{\sum_{j=1}^k \mu_j K_j \;:\; \mu_j \ge 0,\; \sum_{j=1}^k \mu_j = 1\right\}$. This is the unconstrained function class. In applications, one constrains the function class in some way. The class considered in Lanckriet et al. [2004] is

$$\mathcal{F}_{\mathcal{K}_c^+} = \left\{ \mathbf{x} \mapsto \sum_{i=1}^n \alpha_i K(\mathbf{x}_i, \cdot) \;:\; K = \sum_{j=1}^k \mu_j K_j,\; \mu_j \ge 0,\; \sum_{j=1}^k \mu_j = 1,\; \boldsymbol{\alpha}^\top K(\mathcal{T})\boldsymbol{\alpha} \le 1/\gamma^2 \right\} \qquad (3)$$

where $\gamma > 0$ is a margin parameter and $K(\mathcal{T})_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$ is the Gram matrix of K on the dataset $\mathcal{T}$.

Theorem 23 *(Kernel learning) Consider the class $\mathcal{F}_{\mathcal{K}_c^+}$ defined in Eq. (3). Let $K_j(\mathbf{x}, \mathbf{x}) \le B$ for $1 \le j \le k$ and $\mathbf{x} \in \mathcal{X}$. Then, $\mathcal{R}_\mathcal{T}(\mathcal{F}_{\mathcal{K}_c^+}) \le e\sqrt{\frac{B \log k}{\gamma^2 n}}$.*

The proof follows directly from the equivalence between kernel learning and group Lasso Bach [2008], and then applying our bound on the class $\mathcal{W}_{2,1}$. For completeness, we give a rigorous proof in the appendix.

Note that the dependence on the number of base kernels, k, is rather mild (only logarithmic) — implying that we can learn a kernel as a (convex) combination of a rather large number of base kernels. Also, let us discuss how the above improves upon the prior bounds provided by Lanckriet et al. [2004] and Srebro and Ben-David [2006] (neither of which had logarithmic k dependence). The former proves a bound of $O\left(\sqrt{\frac{Bk}{\gamma^2 n}}\right)$ which is quite inferior to our bound. We cannot compare our bound directly to the bound in Srebro and Ben-David [2006] as they do not work with Rademacher complexities. However, if one compares the resulting generalization error bounds, then their bound is $O\left(\sqrt{\frac{k \log \frac{n^3 B}{\gamma^2 k} + \frac{B}{\gamma^2} \log \frac{\gamma n}{\sqrt{B}} \log \frac{nB}{\gamma^2}}{n}}\right)$ and ours is $O\left(\sqrt{\frac{B \log k}{\gamma^2 n}}\right)$. If $k \ge n$, their bound is vacuous (while ours is still meaningful). If $k \le n$, our bound is better.

Finally, we note that recently Ying and Campbell [2009] devoted a dedicated effort to derive a result similar to Theorem. 23 using a Rademacher chaos process of order two over candidate kernels. In contrast to their proof, our result seamlessly follows from the general framework of deriving bounds using the strong-convexity/strong-smoothness duality.

Acknowledgements

We thank Andreas Argyriou, Shmuel Friedland & Karthik Sridharan for helpful discussions.

References

Alekh Agarwal, Alexander Rakhlin, and Peter Bartlett. Matrix regularization techniques for online multitask learning. Technical report, EECS Department, University of California, Berkeley, 2008.

Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

Francis Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 9, 2008.

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Invent. Math.*, 115:463–482, 1994.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

R. Bhatia. *Matrix Analysis*. Springer, 1997.

J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, pages 251–262, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000.

C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

A. J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

A. Juditsky and A. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *submitted to Annals of Probability*, 2008.

S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 22*, 2008.

J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Journal of Machine Learning*, 45(3):301–329, July 2001.

J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, January 1997.

G.R.G. Lanckriet, N. Cristianini, P.L. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(2):173–183, 1995.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468*, Mar 2009.

Andreas Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 2006.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

A.Y. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

G. Obozinski, B. Taskar, and M Jordan. Joint covariate selection for grouped classification. Technical Report 743, Dept. of Statistics, University of California Berkeley, 2007.

I. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *Ann. Probab*, 22(4):1679–1706, 1994.

G. Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20(3–4):326–350, 1975.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

S. Shalev-Shwartz and Y. Singer. Convex repeated games and Fenchel duality. In *Advances in Neural Information Processing Systems 20*, 2006.

S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning Journal*, 2007.

S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2008.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, pages 169–183, 2006.

M. Warmuth and D. Kuzmin. Online variance minimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *Advances in Neural Information Processing Systems 23*, 2009.

Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In *COLT*, 2009.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

C. Zalinescu. *Convex analysis in general vector spaces*. World Scientific Publishing Co. Inc., River Edge, NJ, 2002.

A Convex Analysis and Matrix Computation

A.1 Convex analysis

We briefly recall some key definitions from convex analysis that are useful throughout the paper (for details, see any of the several excellent references on the subject, e.g. Borwein and Lewis [2006], Rockafellar [1970]). We consider convex functions $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, where $\mathcal{X}$ is a Euclidean vector space equipped with an inner product $\langle \cdot, \cdot \rangle$. We denote $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$. Recall that the subdifferential of f at $x \in \mathcal{X}$, denoted by $\partial f(x)$, is defined as $\partial f(x) := \{y \in \mathcal{X} \ : \ \forall z,\ f(x+z) \geq f(x) + \langle y, z \rangle\}$. The Fenchel conjugate $f^\star : \mathcal{X} \to \mathbb{R}^*$ is defined as $f^\star(y) := \sup_{x \in \mathcal{X}} \langle x, y \rangle - f(x)$.

We also deal with a variety of norms in this paper. Recall that given a norm $\| \cdot \|$ on $\mathcal{X}$, its dual norm is defined as $\|y\|_\star := \sup\{\langle x, y \rangle \ : \ \|x\| \leq 1\}$. An important property of the dual norm is that the Fenchel conjugate of the function $\frac{1}{2}\|x\|^2$ is $\frac{1}{2}\|y\|_\star^2$.

The definition of Fenchel conjugate implies that for any x, y, $f(x) + f^\star(y) \geq \langle x, y \rangle$, which is known as the Fenchel-Young inequality. An equivalent and useful definition of the subdifferential can be given in terms of the Fenchel conjugate: $\partial f(x) = \{y \in \mathcal{X} \ : \ f(x) + f^*(y) = \langle x, y \rangle\}$.

A.2 Convex analysis of matrix functions

We consider the vector space $\mathcal{X} = \mathbb{R}^{m \times n}$ of real matrices of size $m \times n$ and the vector space $\mathcal{X} = \mathbb{S}^n$ of symmetric matrices of size $n \times n$, both equipped with the inner product, $\langle \mathbf{X}, \mathbf{Y} \rangle := \mathrm{Tr}(\mathbf{X}^\top \mathbf{Y})$. Recall that any matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ can be decomposed as $\mathbf{X} = \mathbf{U}\mathrm{Diag}(\sigma(\mathbf{X}))\mathbf{V}$ where $\sigma(\mathbf{X})$ denotes the vector $(\sigma_1, \sigma_2, \ldots \sigma_l)$ ($l = \min\{m, n\}$), where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_l \geq 0$ are the singular values of $\mathbf{X}$ arranged in non-increasing order, and $\mathbf{U} \in \mathbb{R}^{m \times m}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices. Also, any matrix $\mathbf{X} \in \mathbb{S}^n$ can be decomposed as, $X = \mathbf{U}\mathrm{Diag}(\lambda(\mathbf{X}))\mathbf{U}^\top$ where $\lambda(\mathbf{X}) = (\lambda_1, \lambda_2, \ldots \lambda_n)$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of $\mathbf{X}$ arranged in non-increasing order, and $\mathbf{U}$ is an orthogonal matrix. Two important results relate matrix inner products to inner products between singular (and eigen-) values

Theorem 24 (von Neumann) *Any two matrices $X, Y \in \mathbb{R}^{m \times n}$ satisfy the inequality*

$$\langle \mathbf{X}, \mathbf{Y} \rangle \leq \langle \sigma(\mathbf{X}), \sigma(\mathbf{Y}) \rangle \ .$$

Equality holds above, if and only if, there exist orthogonal $\mathbf{U}, \mathbf{V}$ such that

$$\mathbf{X} = \mathbf{U}\mathrm{Diag}(\sigma(\mathbf{X}))\mathbf{V} \qquad\qquad \mathbf{Y} = \mathbf{U}\mathrm{Diag}(\sigma(\mathbf{Y}))\mathbf{V} \ .$$

Theorem 25 (Fan) *Any two matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^n$ satisfy the inequality*

$$\langle \mathbf{X}, \mathbf{Y} \rangle \leq \langle \lambda(\mathbf{X}), \lambda(\mathbf{Y}) \rangle \ .$$

Equality holds above, if and only if, there exists orthogonal $\mathbf{U}$ such that

$$\mathbf{X} = \mathbf{U}\mathrm{Diag}(\lambda(\mathbf{X}))\mathbf{U}^\top \qquad\qquad \mathbf{Y} = \mathbf{U}\mathrm{Diag}(\lambda(\mathbf{Y}))\mathbf{U}^\top \ .$$

We say that a function $g : \mathbb{R}^n \to \mathbb{R}^*$ is symmetric if $g(x)$ is invariant under arbitrary permutations of the components of x. We say g is absolutely symmetric if $g(x)$ is invariant under arbitrary permutations and sign changes of the components of x.

Given a function $f : \mathbb{R}^l \to \mathbb{R}^*$, we can define a function $f \circ \sigma : \mathbb{R}^{m \times n} \to \mathbb{R}^*$ as,

$$(f \circ \sigma)(\mathbf{X}) := f(\sigma(\mathbf{X})) \ .$$

Similarly, given a function $g : \mathbb{R}^n \to \mathbb{R}^*$, we can define a function $g \circ \lambda : \mathbb{S}^n \to \mathbb{R}^*$ as,

$$(g \circ \lambda)(\mathbf{X}) := g(\lambda(\mathbf{X})) \ .$$

This allows us to define functions over matrices starting from functions over vectors. Note that when we use $f \circ \sigma$ we are assuming that $\mathcal{X} = \mathbb{R}^{m \times n}$ and for $g \circ \lambda$ we have $\mathcal{X} = \mathbb{S}^n$. The following result allows us to immediately compute the conjugate of $f \circ \sigma$ and $g \circ \lambda$ in terms of the conjugates of f and g respectively.

Theorem 26 *(Lewis [1995]) Let $f : \mathbb{R}^l \to \mathbb{R}^*$ be an absolutely symmetric function. Then,*

$$(f \circ \sigma)^\star = f^\star \circ \sigma \ .$$

Let $g : \mathbb{R}^n \to \mathbb{R}^*$ be a symmetric function. Then,

$$(g \circ \lambda)^\star = g^\star \circ \lambda \ .$$

Proof: Lewis [1995] proves this for singular values. For the eigenvalue case, the proof is entirely analogous to that in Lewis [1995], except that Fan's inequality is used instead of von Neumann's inequality. ■

Using this general result, we are able to define certain matrix norms.

Corollary 27 *(Matrix norms) Let $f : \mathbb{R}^l \to \mathbb{R}^*$ be absolutely symmetric. Then if $f = \|\cdot\|$ is a norm on $\mathbb{R}^l$ then $f \circ \sigma = \|\sigma(\cdot)\|$ is a norm on $\mathbb{R}^{m\times n}$. Further, the dual of this norm is $\|\sigma(\cdot)\|_\star$.*

Let $g : \mathbb{R}^n \to \mathbb{R}^*$ be symmetric. Then if $g = \|\cdot\|$ is a norm on $\mathbb{R}^n$ then $g \circ \lambda = \|\lambda(\cdot)\|$ is a norm on $\mathbb{S}^n$. Further, the dual of this norm is $\|\lambda(\cdot)\|_\star$.

Another nice result allows us to compute subdifferentials of $f \circ \sigma$ and $g \circ \lambda$ (note that elements in the subdifferential of $f \circ \sigma$ and $g \circ \lambda$ are matrices) from the subdifferentials of f and g respectively.

Theorem 28 *(Lewis [1995]) Let $f : \mathbb{R}^l \to \mathbb{R}^*$ be absolutely symmetric and $\mathbf{X} \in \mathbb{R}^{m\times n}$. Then,*

$$\partial(f \circ \sigma)(\mathbf{X}) = \{\mathbf{U}\mathrm{Diag}(\mu)\mathbf{V}^\top \ : \ \mu \in \partial f(\sigma(\mathbf{X}))\mathbf{U}, \mathbf{V} \text{ orthogonal}, \ \mathbf{X} = \mathbf{U}\mathrm{Diag}(\sigma(\mathbf{X}))\mathbf{V}^\top\}$$

Let $g : \mathbb{R}^n \to \mathbb{R}^*$ be symmetric and $\mathbf{X} \in \mathbb{S}^n$. Then,

$$\partial(g \circ \lambda)(\mathbf{X}) = \{\mathbf{U}\mathrm{Diag}(\mu)\mathbf{U}^\top \ : \ \mu \in \partial g(\lambda(\mathbf{X}))\mathbf{U} \text{ orthogonal}, \ \mathbf{X} = \mathbf{U}\mathrm{Diag}(\lambda(X))\mathbf{U}^\top\}$$

Proof: Again, Lewis [1995] proves the case for singular values. For the eigenvalue case, again, the proof is identical to that in Lewis [1995], except that Fan's inequality is used instead of von Neumann's inequality. ■

B Technical Proofs

B.1 Proof of Theorem. 3

First, [Shalev-Shwartz, 2007, Lemma 15] yields one half of the claim (f strongly convex $\Rightarrow f^\star$ strongly smooth). It is left to prove that f is strongly convex assuming that $f^\star$ is strongly smooth. For simplicity assume that $\beta = 1$. Denote $g(y) = f^\star(x+y) - (f^\star(x) + \langle \nabla f^\star(x), y\rangle)$. By the smoothness assumption, $g(y) \le \frac{1}{2}\|y\|_\star^2$. This implies that $g^\star(a) \ge \frac{1}{2}\|a\|^2$ because of [Shalev-Shwartz and Singer, 2008, Lemma 19] and that the conjugate of half squared norm is half squared of the dual norm. Using the definition of g we have

$$\begin{aligned}
g^\star(a) &= \sup_y \langle y, a\rangle - g(y) \\
&= \sup_y \langle y, a\rangle - (f^\star(x+y) - (f^\star(x) + \langle \nabla f^\star(x), y\rangle)) \\
&= \sup_y \langle y, a + \nabla f^\star(x)\rangle - f^\star(x+y) + f^\star(x) \\
&= \sup_z \langle z - x, a + \nabla f^\star(x)\rangle - f^\star(z) + f^\star(x) \\
&= f(a + \nabla f^\star(x)) + f^\star(x) - \langle x, a + \nabla f^\star(x)\rangle
\end{aligned}$$

where we have used that $f^{\star\star} = f$, in the last step. Denote $u = \nabla f^\star(x)$. From the equality in Fenchel-Young (e.g. [Shalev-Shwartz and Singer, 2008, Lemma 17]) we obtain that $\langle x, u\rangle = f^\star(x) + f(u)$ and thus

$$g^\star(a) = f(a+u) - f(u) - \langle x, a\rangle \ .$$

Combining with $g^\star(a) \ge \frac{1}{2}\|a\|^2$, we have

$$f(a+u) - f(u) - \langle x, a\rangle \ \ge \ \frac{1}{2}\|a\|^2 \ , \tag{4}$$

which holds for all a, x, with $u = \nabla f^\star(x)$.

Now let us prove that for any point u' in the relative interior of the domain of f that if $x \in \partial f(u')$ then $u' = \nabla f^\star(x)$. Let $u := \nabla f^\star(x)$ and we must show that $u' = u$. By Fenchel-Young, we have that $\langle x, u'\rangle = f^\star(x) + f(u')$, and, again by Fenchel-Young (and $f^{\star\star} = f$), we have $\langle x, u\rangle = f^\star(x) + f(u)$. We can now apply Equation Eq. (4), to obtain:

$$\begin{aligned}
0 &= \langle x, u\rangle - f(u) - (\langle x, u'\rangle - f(u')) \\
&= f(u') - f(u) - \langle x, u' - u\rangle \ \ge \ \frac{1}{2}\|u' - u\|^2 \ ,
\end{aligned}$$

which implies that $u' = \nabla f^\star(x)$.

Next, let u_1, u_2 be two points in the relative interior of the domain of f, let $\alpha \in (0,1)$, and let $u = \alpha u_1 + (1-\alpha)u_2$. Let $x \in \partial f(u)$ (which is non-empty [1]). We have that $u = \nabla f^\star(x)$, by the previous

[1] The set $\partial f(u)$ is not empty for all u in the relative interior of the domain of f. See the relative max formula in [Borwein and Lewis, 2006, page 42] or [Rockafellar, 1970, page 253]. If u is not in the interior of f, then $\partial f(u)$ is empty. But, a function is defined to be essentially strictly convex if it is strictly convex on any subset of $\{u : \partial f(u) \neq \emptyset\}$. The last set is called the domain of ∂f and it contains the relative interior of the domain of f, so we're ok here.

argument. Now we are able to apply Equation Eq. (4) twice, once with $a = u_1 - u$ and once with $a = u_2 - u$ (and both with x) to obtain

$$\begin{aligned} f(u_1) - f(u) - \langle x, u_1 - u\rangle &\geq \frac{1}{2}\|u_1 - u\|^2 \\ f(u_2) - f(u) - \langle x, u_2 - u\rangle &\geq \frac{1}{2}\|u_2 - u\|^2 \end{aligned}$$

Finally, summing up the above two equations with coefficients α and $1-\alpha$ we obtain that f is strongly convex.

B.2 Proof of Theorem. 12

Note that an equivalent definition of σ-smoothness of a function f w.r.t. a norm $\|\cdot\|$ is that, for all x, y and $\alpha \in [0,1]$, we have

$$f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y) - \frac{1}{2}\sigma\alpha(1-\alpha)\|x-y\|^2 .$$

Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ be arbitrary matrices with columns $\mathbf{X}^i$ and $\mathbf{Y}^i$ respectively. We need to prove

$$\|(1-\alpha)\mathbf{X} + \alpha\mathbf{Y}\|^2_{\Psi,\Phi} \geq \alpha\|\mathbf{X}\|^2_{\Psi,\Phi} + (1-\alpha)\|\mathbf{Y}\|^2_{\Psi,\Phi} - \frac{1}{2}(\sigma_1+\sigma_2)\alpha(1-\alpha)\|\mathbf{X}-\mathbf{Y}\|^2_{\Psi,\Phi} . \quad (5)$$

Using smoothness of Ψ and that Φ is a Q-norm, we have,

$$\begin{aligned} \|(1-\alpha)\mathbf{X} + \alpha\mathbf{Y}\|^2_{\Psi,\Phi} &= (\Phi^2 \circ \sqrt{})(\ldots, \Psi^2(\alpha\mathbf{X}^i + (1-\alpha)\mathbf{Y}^i), \ldots) \\ &\geq (\Phi^2 \circ \sqrt{})(\ldots, \alpha\Psi^2(\mathbf{X}^i) + (1-\alpha)\Psi^2(\mathbf{Y}^i) - \frac{1}{2}\sigma_1\alpha(1-\alpha)\Psi^2(\mathbf{X}^i - \mathbf{Y}^i), \ldots) \\ &\geq (\Phi^2 \circ \sqrt{})(\ldots, \alpha\Psi^2(\mathbf{X}^i) + (1-\alpha)\Psi^2(\mathbf{Y}^i), \ldots) \\ &\quad - \frac{1}{2}\sigma_1\alpha(1-\alpha)(\Phi^2 \circ \sqrt{})(\ldots, \Psi^2(\mathbf{X}^i - \mathbf{Y}^i), \ldots) \\ &= \Phi^2(\ldots, \sqrt{\alpha\Psi^2(\mathbf{X}^i) + (1-\alpha)\Psi^2(\mathbf{Y}^i)}, \ldots) - \frac{1}{2}\sigma_1\alpha(1-\alpha)\|\mathbf{X}-\mathbf{Y}\|^2_{\Psi,\Phi} . \quad (6) \end{aligned}$$

Now, we use that, for any $x, y \geq 0$ and $\alpha \in [0,1]$, we have $\sqrt{\alpha x^2 + (1-\alpha)y^2} \geq \alpha x + (1-\alpha)y$. Thus, we have

$$\begin{aligned} &\Phi^2(\ldots, \sqrt{\alpha\Psi^2(\mathbf{X}^i) + (1-\alpha)\Psi^2(\mathbf{Y}^i)}, \ldots) \\ &\geq \Phi^2(\ldots, \alpha\Psi(\mathbf{X}^i) + (1-\alpha)\Psi(\mathbf{Y}^i), \ldots) \\ &\geq \alpha\Phi^2(\ldots, \Psi(\mathbf{X}^i), \ldots) + (1-\alpha)\Phi^2(\ldots, \Psi(\mathbf{Y}^i), \ldots) \\ &\quad - \frac{1}{2}\sigma_2\alpha(1-\alpha)\Phi^2(\ldots, \Psi(\mathbf{X}^i) - \Psi(\mathbf{Y}^i), \ldots) \\ &\geq \alpha\|\mathbf{X}\|^2_{\Psi,\Phi} + (1-\alpha)\|\mathbf{Y}\|^2_{\Psi,\Phi} - \frac{1}{2}\sigma_2\alpha(1-\alpha)\Phi^2(\ldots, \Psi(\mathbf{X}^i - \mathbf{Y}^i), \ldots) \\ &= \alpha\|\mathbf{X}\|^2_{\Psi,\Phi} + (1-\alpha)\|\mathbf{Y}\|^2_{\Psi,\Phi} - \frac{1}{2}\sigma_2\alpha(1-\alpha)\|\mathbf{X}-\mathbf{Y}\|^2_{\Psi,\Phi} \end{aligned}$$

Plugging this into Eq. (6) proves Eq. (5).

B.3 Proof of Theorem. 23

Let $\mathcal{H}_j$ be the RKHS of K_j, $\mathcal{H}_j = \left\{\sum_{i=1}^l \alpha_i K_j(\tilde{\mathbf{x}}_i, \cdot) \;:\; l > 0,\; \tilde{\mathbf{x}}_i \in \mathcal{X},\; \boldsymbol{\alpha} \in \mathbb{R}^l\right\}$ equipped with the inner product

$$\left\langle \sum_{i=1}^l \alpha_i K_j(\tilde{\mathbf{x}}_i, \cdot), \sum_{j=1}^m \alpha'_i K_j(\tilde{\mathbf{x}}'_j, \cdot) \right\rangle_{\mathcal{H}_j} = \sum_{i,j} \alpha_i\alpha'_j K_j(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_j)$$

Consider the space $\mathcal{H} = \mathcal{H}_1 \times \ldots \times \mathcal{H}_k$ equipped with the inner product $\langle \vec{u}, \vec{v}\rangle := \sum_{i=1}^k \langle u_i, v_i\rangle_{\mathcal{H}_i}$. For $\vec{w} \in \mathcal{H}$, let $\|\cdot\|_{2,1}$ be the norm defined by $\|\vec{w}\|_{2,1} = \sum_{i=1}^k \|w_i\|_{\mathcal{H}_i}$. It is easy to verify that $\mathcal{F}_{\mathcal{K}_c^+} \subseteq \mathcal{F}_r$ where $\mathcal{F}_r := \{\mathbf{x} \mapsto \left\langle \vec{w}, \vec{\phi}(\mathbf{x})\right\rangle \;:\; \vec{w} \in \mathcal{H},\; \|\vec{w}\|_{2,1} \leq 1/\gamma\}$, and $\vec{\phi}(\mathbf{x}) = (K_1(\mathbf{x}, \cdot), \ldots, K_k(\mathbf{x}, \cdot)) \in \mathcal{H}$. Since $\|K_j(\mathbf{x}, \cdot)\|_{\mathcal{H}_j} \leq \sqrt{B}$, we also have $\|\vec{\phi}(\mathbf{x})\|_{2,s} \leq k^{1/s}\sqrt{B}$ for any $\mathbf{x} \in \mathcal{X}$. The claim now follows directly from the results we derived in Section 2.